The Merck Challenge

The Merck Challenge is a competition organized by Merck research group and the Department of Statistics at Rutgers as a training component of FSRM and MSDS programs in the Department of Statistics at Rutgers.

The Merck research group has identified three important and challenging problems:

- Evaluation of Hyperparameter Tunability in Statistical Machine
- Evaluation of Random Projections in Quantitative Structure-Activity Relationship (QSAR) Models
- R/Shiny Application for Loading, Processing and Storing Biological Assay Data

All of the three problems are motivated by the real practice in drug development and the solutions may have big impacts. The detailed description of the projects can be found at the end of this announcement.

Challenge Time: May 27th – August 19th (12 weeks)

Challenge Committee: There is a Challenge Committee formed by both researchers in Merck and faculties in the Department of Statistics at Rutgers. The committee will take in charge of selecting participation teams, monitoring the project progress and evaluating the team performance.

Challenge Format: You will work on the project either by yourself or in a team with up to three team members. The team needs to pick up one of the three projects to work on. You can form the team by yourselves. If you need help to find team members, please contact Mr. Mohannad Aama (mohannad.aama@rutgers.edu) and Sijian Wang (sijian.wang@stat.rutgers.edu). We can help you match with other students with the same need.

Challenge Application: If you are interested in participating in the challenge, please send one email with the names and resumes of all your team members as well as the name of the selected project to mohannad.aama@rutgers.edu and sijian.wang@stat.rutgers.edu. The Challenge Committee will select participation teams. **The application deadline is May 13th 2020**.

Challenge Award: The teams' performance will be evaluated by the Challenge Committee. There will be $1,000 reward in total for winning teams.

Project Format: Each team will have one or two supervisors from Merck research group. The teams work on the same project will have an all-together weekly group meeting with project supervisors.

To monitor your progress, each team is required to maintain a google document to record and track your work in the following format.

Time	To-do during this week	Expected results in the coming week
Week 1	1.	1.

	2. 3.	2. 3.
Week 2	……	……

The document should be updated in the same day when the weekly group meeting is held. The document will be viewable only by yourself and the committee. Other teams will not be able to view your document.

There will be two interim oral reports given on June 24th and July 22nd (4 weeks and 8 weeks after the start of the project). Your progress will be evaluated by The Challenge Committee. If your performance is not satisfied, you will not be allowed to continue in the challenge. The final report will be on August 19th. An oral presentation as well as a written report should be delivered.

A word to the participants: The rules of this competition are set by the competition hosts (the challenge committee). This fact must be acknowledged by participants entering the competition.

Data Science Projects with Rutgers

Evaluation of Hyperparameter Tunability in Statistical Machine Learning
Richard Baumgartner & Andy Liaw

Understanding of hyperparameter tuning for machine learning algorithms is essential for their appropriate application in real world problems. This project will review current methods for hyperparameter tuning for frequently used machine learning algorithms, carry out evaluations using simulation and real-world data and develop a guidance for applications. The main reference for this work is Probst et al., 2019.

The data science project comprises of following goals and milestones:

1) Review of current methods and available R and/or Python packages for tuning of the machine learning algorithms including Bayesian hyperparameter optimization
2) Design of a systematic simulation study to evaluate the hyperparameter optimization of frequently used algorithms. These will include xgboost, different incarnations of random forests and neural networks, etc.
3) Evaluation of the parameter tuning techniques on real life data sets
4) Development of recommendations on the hyperparameter tunability of the algorithms investigated

Reference

Probst at al. Tunability: Importance of hyperparameters of machine learning algorithms. Journal of Machine Learning Research 20 (2019) 1-32

Evaluation of Random Projections in Quantitative Structure-Activity Relationship (QSAR) Models

Andy Liaw

Random projections (RP) has been touted as a highly efficient method for dimension reduction. This project will review current results on RP, and evaluate its usefulness as a pre-processing step to the supervised learning task in Quantitative Structure-Activity Relationship (QSAR) models. Specifically, what if any efficiency (in terms of computational time and memory requirement) can be achieved, as well as impact on prediction performance. This can be evaluated on ML methods that are sensitive to high dimensional inputs or those with very high computational costs.

The data science project comprises of following goals and milestones:

1) Review of current methods and relevant results for RP and available R and/or Python packages.
2) Design of a study to evaluate how RP impact prediction performance of ML methods, and the potential gain in computational efficiency. ML methods to consider may include Deep Neural Networks, Support Vector Machines, Bayesian Additive Regression Trees (BART), etc.
3) Merck's published benchmark QSAR datasets can be used. These can be found as supplementary information to Ma et al (2015): https://pubs.acs.org/doi/10.1021/ci500747n
4) Stretch goal: Find ways to map variable importance measures back to original variables.

Reference

Ma et al. "Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships" *Journal of Chemical Information and Modeling*, 2019, 55 (2), 263-274.

R/Shiny Application for Loading, Processing and Storing Biological Assay Data

Tom Steinmetz, Jennifer Nguyen

Many instruments used in biological assays generate data in a simple text format. The text output includes individual readings and basic meta data but does not include critical information such as sample identifiers or sample content information. The goal of this project is to create an R/Shiny application that performs complete processing of instrument raw data files, maps the readings in the data files to sample identifiers and retains the data for future reference. The use of an R/Shiny interface will allow the application to be used by laboratory scientists unfamiliar with the R programming language. The application will be required to upload files, provide input forms for users to enter data related to the file, process the inputs and finally store the results for future reference.

The data science project comprises of following goals and milestones:

1) Create an R/Shiny user interface that upload and process a text from an assay instrument. The processing will include extracting readings and basic meta data from the file.
2) The R/Shiny application will be updated to include input forms for users to enter critical information such as sample identifiers and sample content information.
3) Include an output summary of the processed data and allow users to download the processed data in a standard format such as Excel.
4) Create a historical saving feature to allow users to access previously processed data including the original raw data file.
5) Stretch goal 1: created graphical trends for historical data uploaded to the application.
6) Stretchgoal 2: Update application to process more advanced multiplexed data files. Multiplexed data files report multiple readings per sample.

Reference

R packages: shiny, shinydashboard, openxlsx, rhandsontable, ggplot2